Title

Sequence Based Indexing and Retrieval Method for Text Documents

Background of the Present Invention

Field of Invention

The present invention relates to a database search engine and, more particularly, to a sequence based indexing and retrieval method for a collection of text documents, which is adapted to produce a ranked list of the text documents relative to a user's query by matching representative token sequences of each document in the collection against the token sequence of the query.

Description of Related Arts

The main task of a text retrieval system is to help the user find, from a collection of text documents, those that are relevant to his query. The system usually creates an index for the text collection to accelerate the search process. Inverted indices (files) are a popular way for such indexing. For each token (word or character), the index records the identifier of every document containing the token. Some extension of inverted indices records not only which documents contain a particular token, but also the positions where in a document the token appears.

Traditional text retrieval models (such as the boolean model and the vector model) are only concerned with the existence of a token in the target document and are insensitive to token order or position. Given a query "United Nations," a traditional retrieval system would consider a document with both "United" and "Nation" (after stemming) as equally relevant as a document that actually contains the phrase "United Nations." One solution to this problem is to index phrases, which would considerably increase the size of the index and require the use of a dictionary. An alternative is for a retrieval system to utilize positional information. If the system takes positional information into account, a document that contains "United" and "Nations" in

consecutive positions will be ranked higher than a document with both words in separate positions. The present invention exploits positional information to its fullest potential.

Summary of the Present Invention

A main object of the present invention is to provide a sequence based indexing and retrieval method for a collection of text documents, which treats the documents and queries as sequences of token-position pairs and estimates the similarity between the document and query, so as to enhance the retrieval effectiveness while performing the query on the text documents.

Another object of the present invention is to provide a sequence based indexing and retrieval method for a collection of text documents, wherein the similarity measurement includes the token appearance, the token order, and the token consecutiveness, such that the approximate matching and fault-tolerant capability are substantially enhanced so as to precisely determine the similarity between the document and query.

Another object of the present invention is to provide a sequence based indexing and retrieval method for a collection of text documents, wherein the text document is pre-processed to select the candidate document therefrom to match with the query token sequence so as to enhance the speed of the retrieval process.

Another object of the present invention is to provide a sequence based indexing and retrieval method for a collection of text documents, wherein each of the text documents is indexed to measure a differentiating position of each two adjacent document tokens in the text document so as to enhance the process of matching the query token sequence with the document token sequence.

Another object of the present invention is to provide a sequence based indexing and retrieval method for a collection of text documents, which is specifically designed as a flexible and modular process that is easy to adjust, modify, and add modules or functionalities for further development.

Another object of the present invention is to provide a sequence based indexing and retrieval method for a collection of text documents, which is adapted to process the text document in Chinese, English, numbers, punctuations, and symbols, so as to enhance the practical use of the present invention.

Accordingly, in order to accomplish the above objects, the present invention provides a sequence based indexing and retrieval method for a text document, comprising the steps of:

(a) generating a query token sequence, having at least a query token, from a query submitted by a user;

(b) generating at least a representative token sequence, having at least a document token, from each of said text documents that contain at least one token of said query token sequence;

(c) measuring a similarity between said query token sequence and each of said representative token sequences; and

(d) retrieving said text documents in responsive to said similarity of said representative token sequence with respect to said query token sequence with a ranking order in accordance with a token appearance score, a token order score, and a token consecutiveness score, provided that for a document with two representative token sequences, its similarity is determined by the representative token sequence with a higher score.

The similarity measurement is preformed by determining a token appearance score, a token order score, and a token consecutiveness score of the representative token sequence with respect to the query token sequence. Therefore, the total score of the token appearance, the token order, and the token consecutiveness is determined as a similarity index to illustrate the similarity between the representative token sequence and the query token sequence, so as to precisely and effectively retrieve the text document.

These and other objectives, features, and advantages of the present invention will become apparent from the following detailed description, the accompanying drawings, and the appended claims.

Brief Description of the Drawings

Fig. 1 is a flow chart illustrating a sequence based indexing and retrieval method for a collection of text documents according to a preferred embodiment of the present invention.

Detailed Description of the Preferred Embodiment

Referring to Fig. 1 of the drawings, a sequence based indexing and retrieval method for a text document according to a preferred embodiment of the present invention is illustrated, wherein the method comprises the following steps.

(1) Generate a query token sequence, having at least a query token, from a query submitted by a user.

(2) Generate at least a representative token sequence, having at least a document token, from each of said documents that contain at least one token of said query token sequence.

(3) Measure a similarity between each of the representative token sequences and the query token sequence.

(4) Retrieve the text documents in responsive to said similarity of said representative token sequence with respect to said query token sequence with a ranking order in accordance with a token appearance score, a token order score, and a token consecutiveness score, provided that for a document with two representative token sequences, its similarity is determined by the representative token sequence with a higher score.

In step (1), the query may contain both English and Chinese. A "Tokenizer" process is preformed to transform the query text into the query token sequence. The key of the Tokenizer is its data analysis component. The input data of the data analysis component is text which is represented as a byte array. This component processes the byte array elements one by one. When encountering the first byte of a Chinese character (in BIG5 encoding, the first byte of a Chinese character is range form 'A4' to 'FF'),

combine it with the next byte to construct a Chinese character. When encountering an English letter ('41' to '5A' and '61' to '7A'), the present invention will check the next byte continuously until reaching a non-English and non-hyphen byte. Then, all checked English letters are combined to construct an English word. If we encounter a non-English and non-Chinese byte (for example, numbers), the number will be treated as an independent unit.

After the data analysis component has parsed out a Chinese character, an English word or others, we use the information to construct a new token by its content, type, and position. After we have processed all bytes, a sequence of query tokens will be constructed.

It is worth mentioning that verb patterns vary in the rules of grammar of the English language, such as present tense, past tense, etc, such that the step (1) further comprises a step of stemming the query tokens to encode the text words into the corresponding word stems respectively by a stemmer. For example, the query token "connecting" is encoded to be "connect" as the origin word stem by removing the suffix thereof. However, for some languages, such as Chinese language, the stemming step can be omitted due to the rules of grammar of the language.

After the introduction of the Tokenizer component, we now explain our method. First, we have to build an index for the collection of text documents. For each token, we record not only which documents contain the token but also the positions where in a document the token appears. For example, the index of a token in essence can be expressed as an extended inverted list:

$$((D_1, (P_1, P_2, P_3, \ldots)), (D_2, (P_1, P_2, P_3 \ldots)) \ldots)$$

According to the preferred embodiment, the step (2) further comprises a step of selecting at least a candidate document from the text documents, wherein one of the text documents is selected to be a candidate document when the respective text document contains the at least one token in the query token sequence.

If the query token sequence contains common words, such as "we," the number of possible candidate documents will be large and thus will reduce the efficiency of the retrieval system. The solution is to adopt the "token weights" concept. The basic idea of

this approach is to eliminate tokens with low discrimination power in the query token sequence. Before using this approach, we have to calculate token weights first. We use the inverse document frequency (idf) metric as token weights. With the weight of each token, we can decide a threshold to drop unimportant query tokens in candidate documents selection.

Here we introduce the approach we designed to solve this problem.

1. For a query token sequence, first we will find out the token with highest weight (W_h) and lowest weight (W_l.)

2. A cut-off percentage cp is given by an implementation parameter wherein cp is in the range of between 0 and 1.

3. Check each query token in the query token sequence. If a token's weight is lower than $W_l + cp * (W_h - W_l)$, we determine that the query token is not as important as other query tokens, and does not use it to select candidate documents.

The document token sequence of the text document is obtained as follows: for each token in a query token sequence, the extended inverted list thereof is obtained from the index; and all lists are combined to construct the document token sequences.

After the document token sequence is chosen, we have to find its representative token sequences. A representative token sequence is a segment of the document token sequence. We divide a document token sequence into segments, wherein for each segment, the distance between two adjacent document tokens is no longer than a predetermined positioning value. Two longest segments of the document token sequence are selected as representative token sequences. Here we give an example:

The query token sequence: A_1B_2

The document: AXXBABXXXBAXXXBABABBXXXBA

The given threshold (predetermined positioning value): 3

After the division, we obtain the following four segments: $A_1B_4A_5B_6$, $B_{10}A_{11}$, $B_{15}A_{16}B_{17}A_{18}B_{19}B_{20}$, $B_{24}A_{25}$. The two longest segments, i.e., $A_1B_4A_5B_6$ and $B_{15}A_{16}B_{17}A_{18}B_{19}B_{20}$, will be the representative token sequences of this document.

To summarize, the two longest segments of the document token sequence are selected as representative token sequences wherein the positional differentiation of each adjacent document tokens is no larger than a predetermined positioning value while said corresponding text document is selected as the said candidate document.

The following example mainly illustrates the generation of representative token sequence in form of Chinese language.

The text document is shown as:

Doc # 134

資訊科技日新月異，設計工藝乃至於純藝術也大量運用電腦，來完成人類創造力的美夢，資訊工業策進會二十日起將在資訊科學展示中心舉辦「資訊藝術週」活動，展出時下流行的資訊藝術應用作品。

The query is input as "資策會," wherein the query is transformed into the query token sequence by a Tokenizer as "資$_1$策$_2$會$_3$" while the indices of the relevant document tokens are shown as below:

Extended Inverted Lists:
資 ……,(Doc#134,(1, 41, 54, 65, 81)),(Doc#135,……
策 ……,(Doc#134,(45)),(Doc#135,……
會 ……,(Doc#134,(47)),(Doc#135,……

Reconstruction of the document token sequences (on the basis that the query token sequence is 資$_1$策$_2$會$_3$):

……
Doc#134 資$_1$資$_{41}$策$_{45}$會$_{47}$資$_{54}$資$_{65}$資$_{81}$
Doc#135 ……
……

With a given threshold (a predetermined positioning value) 3, the document token sequence "資$_1$資$_{41}$策$_{45}$會$_{47}$資$_{54}$資$_{65}$資$_{81}$" of Doc#134 is formed into five segments which are "資$_1$, "" 資$_{41}$策$_{45}$會$_{47}$," "資$_{54}$," "資$_{65}$" and "資$_{81}$." Accordingly, the two longest segments of the document token sequences "資$_1$". and "資$_{41}$策$_{45}$會$_{47}$" are selected in this example as representative token sequences for determining the similarity between the between the query token sequence and the document token sequence.

According to the preferred embodiment, the step (3) further comprises the following steps, wherein $D=(d_{i_1},d_{i_2},...,d_{i_j},...,d_{i_m})$ (of m tokens) and $Q=(q_{i_1},q_{i_2},...,q_{i_j},...,q_{i_n})$ (of n tokens) respectively denote the representative token sequence and the query token sequence under similarity measurement.

(3.1) Determine a token appearance (TA) score by measuring a token appearance of the query representative token sequence with respect to the query token sequence.

(3.2) Determine a token order (TO) score by measuring a token order of the representative token sequence with respect to the query token sequence.

(3.3) Determine a token consecutiveness (TC) score by measuring a token consecutiveness of the representative token sequence with respect to the query token sequence.

The step (3.1) comprises the following sub-steps.

(3.1.1) Consult an index of said text documents to determine the weight of each token in the query token sequence.

(3.1.2) Calculate a sum of the weights of the query tokens that appear in the representative token sequence.

(3.1.3) Output a token appearance score of the token appearance by calculating the fraction of the sum divided by the total weight of all query tokens.

As mentioned above, the weight of a query token is measured by (idf + 1). Accordingly, the following equation illustrates the determination of the token appearance TA.

Token Appearance (TA):

$$TA(D,Q) = \frac{\sum_{j=1}^{n} t(q_{i_j}) \times w(q_{i_j})}{\sum_{j=1}^{n} w(q_{i_j})},$$

wherein $w(q_{i_j})$ represents the weight of the "j_{th}" query token.

Accordingly, $t(q_{i_j})$ = 1 if the "j_{th}" query token is shown in the representative token sequence and $t(q_{i_j})$ = 0 if the "j_{th}" query token is not shown in the representative token sequence.

The object of the token order (TO) measurement is to capture the word/character ordering, wherein the step (3.2) comprises the following sub-steps.

(3.2.1) Determine a length of the longest common subsequence of the representative token sequence and the query token sequence;

(3.2.2) Determine a length of the representative token sequence;

(3.2.3) Determine a length of the query token sequence; and

(3.2.4) Output the token order score of said token order by calculating a fraction of the length of the longest common subsequence divided by an average sum of the length of the representative token sequence and the length of the query token sequence.

Accordingly, the equation for the token order TO is:

Token Ordering (TO):

$$TO(D,Q)=\frac{|LCS(D,Q)|}{(|D|+|Q|)\div 2},$$

where LCS(D,Q) is the *longest common subsequence* of D and Q and |S| denotes the length of sequence S.

The object of the token consecutiveness (TC) measurement is to capture the distribution of the query tokens, wherein the step (3.3) further comprises the following sub-steps.

(3.3.1) Determine a relative distance between a positional differentiation of each adjacent document tokens and a positional differentiation of said adjacent document tokens in the query token sequence.

(3.3.2) Output the token consecutiveness score of the token consecutiveness by calculating a fraction of a sum of the inverses of the relative distances divided by the number of pairs of adjacent tokens, which equals the length of the representative token sequence less one.

Token Consecutiveness (TC):

$$TC(D,Q)=\frac{\sum_{j=1}^{m-1}\frac{1}{rd_j}}{m-1},$$

where $rd_j=|(i_{j+1}-i_j)-(pos(d_{i_{j+1}},Q)-pos(d_{i_j},Q))|+1$ where $pos(t_k,Q)$ gives the position of t in Q. When there are more than one possible values for $pos(d_{i_{j+1}},Q)$ or $pos(d_{i_j},Q)$, the values may be chosen such that $|(i_{j+1}-i_j)-(pos(d_{i_{j+1}},Q)-pos(d_{i_j},Q))|$ is as small as possible.

The above three measures all have a score ranging from 0 to 1. A linear combination (weighted sum) of the measures (which also ranges from 0 to 1) can be calculated from $\alpha_1 TA(D,Q)+\alpha_2 TO(D,Q)+\alpha_3 TC(D,Q)$ with a suitable selection of α_1, α_2, and α_3 such that $\alpha_1+\alpha_2+\alpha_3=1$. An implementation may allow the user to select the coefficients.

Therefore, the similarity of the query token sequence is calculated by summing the token appearance score, the token order score, and the token consecutiveness score.

The result shown below illustrates the determination of the similarity between the representative token sequence and the query token sequence.

Following the earlier example, we consider measuring the similarity between the representative token sequence "資$_{41}$策$_{45}$會$_{47}$" and the query token sequence "資$_1$策$_2$會$_3$."

Token appearance TA of the query token sequence:

TA = (1*(1/3)+1*(1/3)+1*(1/3))/(1/3+1/3+1/3)=1

Token order TO of the query token sequence: TO = 3/((3+3)/2)=1

Token consecutiveness TC of the query token sequence: d_1=1+|(45-41)-(2-1)|=4; d_2=1+|(47-45)-(3-2)|=2; TC = ((1/4)+(1/2))/2=0.375

The similarity: 1*1/3 + 1*1/3 + 1*0.375 = 0.792

The following experimental results illustrate the accuracy of the search result by using the present invention in comparison with the bigram method.

Experiment 1 illustrates the query including a person name and the prefix thereof.

Query: 陳總統水扁; wherein "陳水扁" is the name of a person and "總統" is a prefix of the person.

Text Documents	The present invention		Bigram method	
	Point value	Ranking	Point value	Ranking
...陳總統水扁...	1.0	1	1.0	1
...總統陳水扁...	0.861	2	0.5	2
...陳水扁總統...	0.808	3	0.5	2
...陳水扁參選總統...	0.804	4	0.5	2
...陳水扁...	0.654	5	0.25	5
...總統...	0.616	6	0.25	5

Experiment 2 illustrates the query including two person names and a connecting word therebetween.

Query: 辜振甫與汪道涵; wherein "辜振甫" and "汪道涵" are the names of the person and "與" is the connecting word for "辜振甫" and "汪道涵."

Text Documents	The present invention		Bigram method	
	Point value	Ranking	Point value	Ranking
...辜振甫與汪道涵...	1.0	1	1.0	1
...辜振甫與XXX汪道涵...	0.968	2	0.833	2
...辜振甫汪道涵...	0.903	3	0.667	3
...汪道涵與辜振甫...	0.79	4	0.667	3
...汪道涵與XXX辜振甫...	0.787	5	0.667	3
...汪道涵辜振甫...	0.76	6	0.667	3
...辜振甫...	0.614	7	0.333	7
...汪道涵...	0.614	7	0.333	7
...辜汪...	0.33	9	0	9

Experiment 3 illustrates the query including the abbreviation of a noun phrase.

Query	Text Documents	Point Value	
		The Present Invention	Bigram Method
聯合國安理會	...聯合國安全理事會...	0.95	0.6
聯合國安全理事會	...聯合國安理會...	0.789	0.249
臺大	...臺灣大學...	0.875	0
臺灣大學	...臺大...	0.541	0
資策會	...資訊工業策進會...	0.844	0
資訊工業策進會	...資策會...	0.458	0
海基會	...海協交流基金會...	0.844	0
海峽交流基金會	...海基會...	0.458	0
辜汪會談	...辜振甫與汪道涵的會談...	0.875	0.333
辜振甫與汪道涵的會談	...辜汪會談...	0.468	0.111

Therefore, the approximate matching and fault-tolerant capabilities are substantially enhanced so as to efficiently and precisely retrieve text documents with respect to the query submitted by the user.

One skilled in the art will understand that the embodiment of the present invention as shown in the drawings and described above is exemplary only and not intended to be limiting.

It will thus be seen that the objects of the present invention have been fully and effectively accomplished. Its embodiments have been shown and described for the purposes of illustrating the functional and structural principles of the present invention and is subject to change without departure from such principles. Therefore, this invention includes all modifications encompassed within the spirit and scope of the following claims.